DAMSL: Domain Agnostic Meta Score-based Learning

John Cai
Princeton University
jjcai@alumni.princeton.edu

Bill Cai
Massachusetts Institute of Technology
billcai@alum.mit.edu

Shen Sheng Mei
Pensees Pte Ltd
jane.shen@pensees.ai

Abstract

In this paper, we propose Domain Agnostic Meta Score-based Learning (DAMSL), a novel, versatile and highly effective solution that delivers significant out-performance over state-of-the-art methods for cross-domain few-shot learning. We identify key problems in previous meta-learning methods over-fitting to the source domain, and previous transfer-learning methods under-utilizing the structure of the support set. The core idea behind our method is that instead of directly using the scores from a fine-tuned feature encoder, we use these scores to create input coordinates for a domain agnostic metric space. A graph neural network is applied to learn an embedding and relation function over these coordinates to process all information contained in the score distribution of the support set. We test our model on both established CD-FSL benchmarks and new domains and show that our method overcomes the limitations of previous meta-learning and transfer-learning methods to deliver substantial improvements in accuracy across both smaller and larger domain shifts.

1. Introduction

Few-shot learning methods promise to solve one of the most challenging issues in deep learning: the reliance on copious amounts of labelled examples to achieve high accuracies. By doingg so, we can achieve cost savings and accurately classify rare classes of images (e.g. plane crashes) where labelled examples are limited. The problem, however, is that few-shot learning methods fail to perform well when there is a domain-shift. Hence, the practical applications of few-shot learning are severely limited as the few-shot models trained on well-labelled and well-structured research datasets cannot be applied to domains in industry.

The above problem is exacerbated under sharp domain shifts, as shown in the Broader Study of Cross-Domain Few-Shot Learning (BSCD-FSL) [7]. The study found that many few-shot learning methods significantly underperformed compared to transfer-learning as few-shot learning overfitted to the source domain. While transfer-learning methods did perform better, they omitted distributional information contained in the each episode's support set is omitted. This is clearly sub-optimal given the need to maximally use information in the sparse setting [20] [17].

To solve the above issues, we propose Domain Agnostic Meta Score-based Learning (DAMSL). The fundamental idea behind our method is to apply transfer-learning to prevent over-fitting to the source domain, while using metric-learning to exploit the information in each episode's support samples. Furthermore, as metric-learners built on image features are shown to suffer greatly from overfitting to the source domain, we make our metric-learner domain-agnostic by fitting to pre-softmax classification scores from fine-tuned feature encoders.

In our work, we use the BSCD-FSL benchmark [7] and augment it with 4 more test domains for further comparisons. We demonstrate the superiority of our method over existing methods across these 8 distinct test domains, and show a new research direction for score-based boosting in the few-shot classification setting.

2. Relevant Work

Metric-based methods, such as prototypical networks [16], aim to learn a metric function ϕ_m that can be used to classify query images based on their relations to the images in the support set. The key metric-based method that we use is **Graph Neural Network** (GNN) as graph-based convolutions can create more flexible representations [1].

Transfer learning involves reusing features learned from base classes [14], typically by fine-tuning a pre-trained model. A simple extension of fine-tuning would be to **learn to fine-tune**. Methods such as MAML [4] learn an internal representation that can be fine-tuned in a few gradient steps.

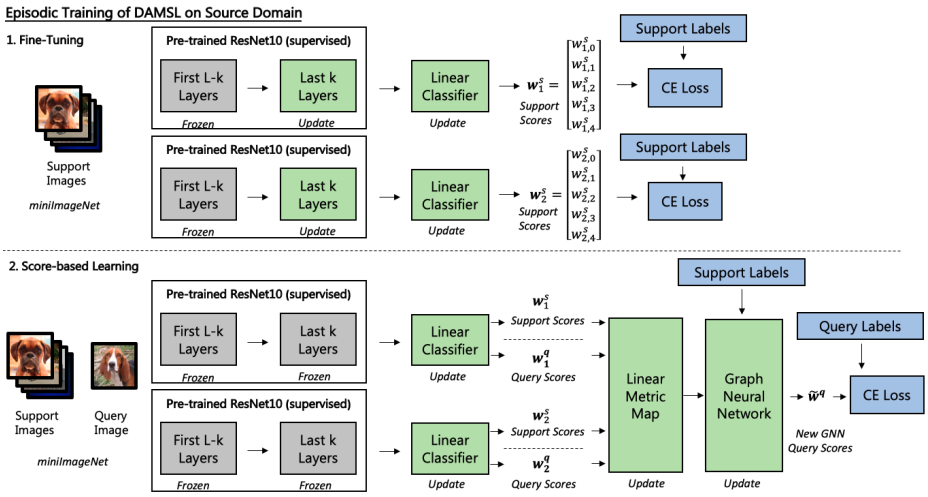

Figure 1. Episodic training on miniImagenet (source domain) for our Proposed DAMSL Model.

3. Methodology

The training process during each episode for our model is shown in Figure 1. On test domains, the same fine-tuning process occurs over the labelled support set, but with gradient updates only within episodes and not between episodes.

3.1. Score-based Metric Learning

Given only a sparse support set from the test domain, it is difficult to precisely fit the feature encoder in a way that neither overfits nor underfits [12]. Creating a hold-out validation set is also prohibitively costly under such conditions.

We begin by fine-tuning a feature vector to obtain $\tilde{\phi}_f(X_s) \in \mathbb{R}^{512}$. Then, we take the linear classifier $\tilde{\phi}_c$ to produce a pre-softmax score vector $\tilde{\phi}_c(\tilde{\phi}_f(X_i)) \in \mathbb{R}^5$, which corresponds to our 5-way classification problem. A typical transfer-learning approach would directly use the score vector for prediction $\hat{Y}_q = \arg\max(\tilde{\phi}_c(\tilde{\phi}_f(X_q))$

Instead, we post-process the score vector by using a metric-learning network ϕ_m. Formally, this gives us: $\hat{Y}_q = \arg\max(\phi_m(Y_s, \tilde{\phi}_c(\tilde{\phi}_f(X_q)), \tilde{\phi}_c(\tilde{\phi}_f(X_s))))$ Thus, we explicitly incorporate information contained in the predictions we can make on the support sets X_s and how these predictions correspond to the labels Y_s. Any biases found in the feature encoder $\tilde{\phi}_f$ and linear classifier $\tilde{\phi}_c$ can be corrected by inferring a distribution from the support set scores, and using that distribution to match the scores of our query samples. This reduces the reliance on the initial fine-tuning process as our decision boundaries from the initial feature encoder are replaced by a metric-based decision boundary constructed from the proximity of the query sample to the support classes. Moreover, the scores form a domain agnostic basis for metric learning because the way that scores relate should not differ significantly across domains.

3.2. Graph Neural Network

The meta-learning module we use is the Graph Neural Network (GNN). We follow the formulation of the GNN for the few-shot problem in [15]. In brief, a GNN acts on local operators of a graph $G = (V, E)$, which for the few-shot learning case is fully connected. A graph convolution layer $GC(.)$ [15] is performed with linear operations on local signals. Formally, we have:

$$GC(S^k) = f\left(\sum_{B \in \mathcal{A}} BS^k \theta^k_{B,q}\right), \; q = d_1, ..., d_{k+1} \quad (1)$$

In the few-shot learning formulation, we can learn the edfe features using the current hidden vertex [15]. We apply a Multi-layer Perceptron (MLP) that takes in the absolute difference between the the output vectors of vertices in the graph [9] [5]. Formally, we have:

$$\tilde{A}^k_{i,j} = \gamma(S^k_i, S^k_j) = MLP(|S^k_i - S^k_j|) \quad (2)$$

These learned edge features are used to propagate information in the graph through the graph convolution in equation 1. Initial vertex features are constructed by taking the score projections and one-hot encoding of labels for the support set or a uniform distribution for the query samples.

3.3. Backbone Variants

We experiment with two variants of the DAMSL model with different feature backbones. In **DAMSL v1**, we have a ResNet10 pre-trained using supervised learning and first-order MAML [13] In **DAMSL v2**, we have two ResNet10 pre-trained using supervised learning but with different optimization strategies - one trained using Adam while other trained using SGD with momentum.

Methods	EuroSAT			CropDisease		
	5-way 5-shot	5-way 20-shot	5-way 50-shot	5-way 5-shot	5-way 20-shot	5-way 50-shot
ProtoNet†	73.29% ± 0.71%	82.27% ± 0.57%	80.48% ± 0.57%	79.72% ± 0.67%	88.15% ± 0.51%	90.81% ± 0.43%
TransFT†	81.76% ± 0.48%	87.97% ± 0.42%	92.00% ± 0.56%	90.64% ± 0.54%	95.91% ± 0.72%	97.48% ± 0.56%
L-Ensem v1	74.64% ± 0.67%	85.52% ± 0.53%	90.38% ± 0.35%	84.65% ± 0.60%	94.40% ± 0.36%	96.89% ± 0.24%
DAMSL v1	85.93% ± 0.68%	95.18% ± 0.35%	97.73% ± 0.25%	95.03% ± 0.42%	99.19% ± 0.14%	99.75% ± 0.08%
L-Ensem v2	81.02% ± 0.62%	90.01% ± 0.37%	93.28% ± 0.30%	90.68% ± 0.51%	97.20% ± 0.26%	98.89% ± 0.16%
DAMSL v2	**90.84% ± 0.54%**	**96.13% ± 0.29%**	**98.15% ± 0.18%**	**97.30% ± 0.32%**	**99.36% ± 0.14%**	**99.73% ± 0.09%**
TransFT (Aug)	82.80% ± 0.60%	90.38% ± 0.67%	93.48% ± 0.57%	92.51% ± 0.84%	97.33% ± 0.43%	98.40% ± 0.41%
L-Ensem v1 (Aug)	77.98% ± 0.66%	89.43% ± 0.39%	93.56% ± 0.31%	88.84% ± 0.54%	97.11% ± 0.23%	98.83% ± 0.18%
FT-GNN v1 (Aug)	82.29% ± 0.63%	92.73% ± 0.63%	93.78% ± 0.63%	94.09% ± 0.46%	98.31% ± 0.35%	98.95% ± 0.25%
S-Proto v1 (Aug)	80.32% ± 0.58%	89.52% ± 0.40%	93.39% ± 0.27%	91.43% ± 0.47%	97.53% ± 0.37%	98.79% ± 0.26%
DAMSL v1 (Aug)	87.30% ± 0.68%	96.53% ± 0.28%	98.37% ± 0.18%	96.01% ± 0.40%	**99.61% ± 0.09%**	99.85% ± 0.06%
L-Ensem v2 (Aug)	83.68% ± 0.54%	91.61% ± 0.34%	94.75% ± 0.25%	92.66% ± 0.45%	98.08% ± 0.20%	99.14% ± 0.11%
DAMSL v2 (Aug)	**91.59% ± 0.49%**	**96.99% ± 0.24%**	**98.60% ± 0.15%**	**97.43% ± 0.31%**	**99.61% ± 0.10%**	**99.87% ± 0.05%**

Methods	ChestX			ISIC		
	5-way 5-shot	5-way 20-shot	5-way 50-shot	5-way 5-shot	5-way 20-shot	5-way 50-shot
ProtoNet†	24.05% ± 1.01%	28.21% ± 1.15%	29.32% ± 1.12%	39.57% ± 0.57%	49.50% ± 0.55%	51.99% ± 0.52%
TransFT†	26.09% ± 0.96%	31.01% ± 0.59%	36.79% ± 0.53%	49.68% ± 0.36%	61.09% ± 0.44%	67.20% ± 0.59%
L-Ensem v1	25.20% ± 0.43%	30.62% ± 0.45%	35.82% ± 0.47%	46.55% ± 0.61%	59.14% ± 0.61%	65.35% ± 0.59%
DAMSL v1	25.99% ± 0.50%	33.47% ± 0.54%	38.37% ± 0.56%	50.68% ± 0.76%	68.58% ± 0.70%	75.55% ± 0.58%
L-Ensem v2	26.38% ± 0.45%	33.46% ± 0.51%	39.81% ± 0.53%	51.93% ± 0.62%	64.21% ± 0.60%	70.28% ± 0.57%
DAMSL v2	**27.22% ± 0.49%**	**35.41% ± 0.56%**	**42.74% ± 0.62%**	**57.35% ± 0.78%**	**70.32% ± 0.70%**	**77.40% ± 0.65%**
TransFT (Aug)	**29.23% ± 0.46%**	36.25% ± 0.55%	40.69% ± 0.56%	51.54% ± 0.64%	62.72% ± 0.62%	69.68% ± 0.59%
L-Ensem v1 (Aug)	26.84% ± 0.44%	34.62% ± 0.48%	40.23% ± 0.56%	48.97% ± 0.65%	62.99% ± 0.60%	70.32% ± 0.57%
FT-GNN v1 (Aug)	26.79% ± 0.50%	35.39% ± 0.60%	35.34% ± 0.54%	52.13% ± 0.84%	65.37% ± 0.73%	62.68% ± 0.65%
S-Proto v1 (Aug)	27.55% ± 0.44%	35.37% ± 0.57%	41.56% ± 0.56%	50.98% ± 0.65%	63.58% ± 0.61%	71.47% ± 0.54%
DAMSL v1 (Aug)	28.08% ± 0.50%	**37.70% ± 0.57%**	**43.04% ± 0.66%**	53.50% ± 0.79%	70.31% ± 0.72%	78.41% ± 0.66%
L-Ensem v2 (Aug)	28.26% ± 0.46%	35.91% ± 0.52%	41.00% ± 0.57%	52.76% ± 0.62%	64.99% ± 0.59%	71.92% ± 0.55%
DAMSL v2 (Aug)	28.86% ± 0.52%	37.04% ± 0.61%	42.87% ± 0.65%	**57.15% ± 0.76%**	**70.87% ± 0.72%**	**78.98% ± 0.62%**

†- as reported in [7]. Aug - with data augmentation during fine-tuning. **Bold** - Best performing in category.

Table 1. Results on BSCD-FSL Benchmark. Includes ablation studies and results from prior work.

4. Results and Discussion

4.1. Experimental Setup

First, we test our model on the BSCD-FSL benchmark. We train on miniImagenet and test on CropDisease [11], EuroSAT [8], ISIC [18] [3] and ChestX [19] (in order of decreasing similarity). CropDisease covers plant disease, EuroSAT covers satellite images, ISIC covers dermoscopic skin lesion images and ChestX covers chest X-ray images.

4 other new datasets are also included: Places [21], Describable Textures Dataset (DTD) [2], CIFAR-100 [10] and Caltech256 [6]. While the images in these other datasets are all natural images, they contain very different types of classification tasks from miniImagenet. DTD is the most distinct, as it requires the model to recognize textures found in many different contexts [2].

4.2. Results on the BS-CDFSL Benchmark

From Table 1, DAMSL with both feature backbones outperform previous methods, even without using data augmentation. We see that DAMSL v1 achieves an average accuracy of **72.13%** while DAMSL v2 achieves an average accuracy of **74.34%**. With data augmentation, DAMSL v1 (Aug) achieves an average accuracy of **74.06%** while DAMSL v2 (Aug) achieves an average accuracy of **74.99%**.

We focus on DAMSL v2 in our comparison with other methods as it has the highest performance. DAMSL v2 (Aug) outperforms TransFT by 6.86% and ProtoNet by 15.21%. From Table 1, we see that DAMSL (Aug) v2 still significantly outperforms TransFT (Aug), with the exception of 5-way 5-shot Chest-X. In terms of average accuracy, DAMSL v2 (Aug) outperforms TransFT (Aug) by 3.84%.

Our method is competitive with typical supervised learning on domains closer to the source domain. Typical supervised learning models for EuroSAT and CropDisease have achieved 98.57% and 99.35% respectively [8] [11]. At our 50-shot results for DAMSL v2 (Aug), we achieve 98.60% and 99.87% respectively on EuroSAT and CropDisease.

4.3. Ablation Studies on BSCD-FSL

To investigate the effect of each component of separately, we include a Linear Ensemble (L-Ensem), a Fine-Tuned en-

Methods	DTD			CIFAR-100		
	5-way 5-shot	5-way 20-shot	5-way 50-shot	5-way 5-shot	5-way 20-shot	5-way 50-shot
TransFT (Aug)	62.17% ± 0.74%	73.49% ± 0.61%	79.25% ± 0.55%	65.89% ± 0.77%	77.60% ± 0.64%	83.64% ± 0.54%
L-Ensem v1 (Aug)	55.80% ± 0.73%	69.20% ± 0.69%	76.39% ± 0.59%	62.07% ± 0.77%	77.85% ± 0.61%	84.17% ± 0.51%
DAMSL v1 (Aug)	58.29% ± 0.89%	77.72% ± 0.71%	85.44% ± 0.58%	67.64% ± 0.93%	86.68% ± 0.64%	93.06% ± 0.42%
L-Ensem v2 (Aug)	60.01% ± 0.74%	73.73% ± 0.65%	79.99% ± 0.55%	66.01% ± 0.42%	80.67% ± 0.59%	86.23% ± 0.46%
DAMSL v2 (Aug)	**68.39% ± 0.89%**	**81.64% ± 0.68%**	**87.14% ± 0.68%**	**76.56% ± 0.85%**	**88.47% ± 0.57%**	**93.92% ± 0.39%**

Methods	Places			Caltech256		
	5-way 5-shot	5-way 20-shot	5-way 50-shot	5-way 5-shot	5-way 20-shot	5-way 50-shot
TransFT (Aug)	67.50% ± 0.75%	76.17% ± 0.67%	80.98% ± 0.57%	75.32% ± 0.70%	84.15% ± 0.58%	88.37% ± 0.47%
L-Ensem v1 (Aug)	70.78% ± 0.92%	74.45% ± 0.65%	79.85% ± 0.57%	70.32% ± 0.72%	83.20% ± 0.55%	87.59% ± 0.48%
DAMSL v1 (Aug)	71.34% ± 0.85%	84.30% ± 0.67%	89.50% ± 0.55%	76.63% ± 0.87%	91.44% ± 0.49%	95.06% ± 0.37%
L-Ensem v2 (Aug)	**75.45% ± 0.80%**	75.87% ± 0.65%	82.13% ± 0.53%	76.31% ± 0.68%	87.59% ± 0.44%	90.93% ± 0.35%
DAMSL v2 (Aug)	75.42% ± 0.80%	**84.74% ± 0.60%**	**90.43% ± 0.51%**	**87.44% ± 0.70%**	**94.53% ± 0.36%**	**96.89% ± 0.24%**

Aug - with data augmentation during fine-tuning. **Bold** - Best performing in category.

Table 2. Results on Additional Test Domains

coder + GNN (FT-GNN), and Score-based ProtoNets (S-Proto) in Table 1. The L-Ensem is a simple addition of the post-softmax scores from the two fine-tuned feature encoders, to see the performance from a simple fine-tuned ensemble. FT-GNN is directly fitted to a feature vector that has been fine-tuned on the support set, to demonstrate the performance boost from score-based learning. The Score-based ProtoNets replaces the GNN module with an embedding MLP and a nearest centroid classifier, to demonstrate the additional gains from a GNN module.

Furthermore, we see that the score-based metric delivers an improvement when used in conjunction with a simple ProtoNets. On all tasks, the S-Proto delivers a better performance compared to L-Ensem. However, it still does not match up to the performance of our proposed DASML model. We attribute this to the GNN's more flexible representations, and the fact that it exploits the full distribution of scores rather than just the mean value of scores. We also demonstrate the value of score-based metric learning as FT-GNN performs substantially worse than DAMSL. This is in line with the expectation that the GNN is trained to interpret domain-specific features, and thus fails on distant domains.

Looking at average accuracy using v1, we observe that L-Ensem yields **69.23%**, FT-GNN yields **69.82%**, S-Proto yields **70.12%**, DAMSL v1 yields **74.06%**. This shows that both parts of DAMSL are most useful when jointly applied.

4.4. Results on other Test Domains

From Table 2, we clearly see that DAMSL v2 delivers out-performance over the previous baselines across almost all settings and all shots, with the only exception of 5-shot setting for Places. In terms of average accuracy, DAMSL v1 achieves **81.48%** while DAMSL v2 achieves **85.46%** . These values are considerably higher ($> 5\%$) than the linear ensembles, which yields **74.31%** and **77.91%** respectively. The results clearly validate our method as DAMSL performs better than previous methods on data-sets other than those in the BSCD-FSL benchmark.

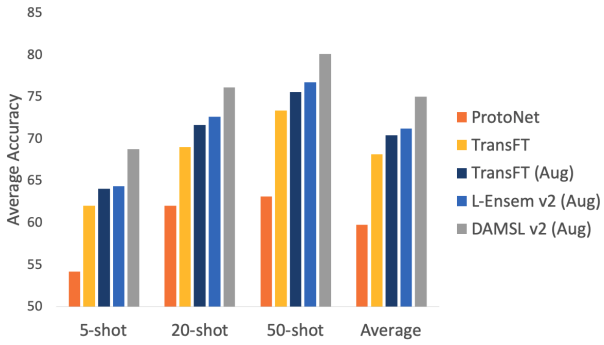

Figure 2. Summary of performance on BSCD-FSL

5. Conclusion

We propose Domain Agnostic Meta Score-based Learning (DAMSL) to address the Cross-Domain Few-Shot Learning problem. On BSCD-FSL, DAMSL v2 achieves **74.99%** accuracy, which significantly outperforms previous best-performing meta-learning and transfer-learning methods by **15.21%** and **6.86%** respectively. From Figure 2, we also see that the performance gains from our method are substantially greater than gains from including data augmentation or adding another feature encoder. Moreover, on the 4 other test domains beyond BSCD-FSL, our method continues to consistently outperform strong baselines.

Ultimately, we not only decisively address the CD-FSL problem, but we also outline a new strand of classification boosting modules that can be attached to any existing model to self-correct initial classification scores by utilizing the distributional information of scores from labelled samples.

References

[1] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 3

[3] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 3

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 1

[5] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272, 2017. 2

[6] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 3

[7] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. ECCV, 2020. 1, 3

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 3

[9] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016. 2

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[11] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016. 3

[12] Akihiro Nakamura and Tatsuya Harada. Revisiting fine-tuning for few-shot learning. *arXiv preprint arXiv:1910.00216*, 2019. 2

[13] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[14] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 1

[15] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. 2

[16] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1

[17] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2252–2262, 2017. 1

[18] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018. 3

[19] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 3

[20] Jian Zhang, Chenglong Zhao, Bingbing Ni, Minghao Xu, and Xiaokang Yang. Variational few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1685–1694, 2019. 1

[21] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 3